

IRLBench: Supplementary Materials

Complete Appendix with Additional Results and Analysis

Khanh-Tung Tran
University College Cork
Cork, Ireland
123128577@umail.ucc.ie

Barry O’Sullivan
University College Cork
Cork, Ireland
b.osullivan@cs.ucc.ie

Duc-Hai Nguyen
University College Cork
Cork, Ireland
125109073@umail.ucc.ie

Hoang D. Nguyen
University College Cork
Cork, Ireland
hn@cs.ucc.ie

Abstract

This document contains the complete supplementary materials for the IRLBench paper, including all prompts used in benchmarking, additional experimental results, statistical analysis, and detailed performance breakdowns. Due to page limitations in the main conference proceedings, these materials are provided separately to ensure full reproducibility and transparency of our work.

ACM Reference Format:

Khanh-Tung Tran, Duc-Hai Nguyen, Barry O’Sullivan, and Hoang D. Nguyen. 2025. IRLBench: Supplementary Materials: Complete Appendix with Additional Results and Analysis. In . ACM, New York, NY, USA, 12 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Prompts Used in Benchmarking

1.1 Response Generation Prompt

The prompt used for all candidate LLMs to generate their responses are shown in Figure 1, including prompting to output their self-assessed confidence score alongside the answer.

1.2 Judgment Prompt

The prompt used for the judge LLM to evaluate candidate responses are shown in Figure 2, inspired from the prompt used in related work. We leverage structured output (e.g., outputting JSON dictionary) to format the LLM’s judgment.

1.3 Vision-Language Model Extraction Prompt

Figure 3 shows the prompt we give to the vision-language model for the extraction task, which includes the expected output format and some illustrative examples.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference’17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/XXXXXXX.XXXXXXX>

2 Examples

The examples for a judge-correct and a judge-incorrect decision are illustrated in Figures 4 and 5, respectively.

3 Other LLMs Performance

We benchmark more LLMs, namely claude-3.7-sonnet and Qwen2.5-VL-32B-Instruct. We plot their results by languages in Figure 6. Their confidences, relatively to their performances, are visualized in Figure 7. Overall, these results provide more support and evidence for our findings in the main paper.

4 Text-only Benchmarking Performance

We benchmark both vision-language models and language-only models on the text subset of IRLBench, comprising of 1392 samples across the 12 subjects (81.89% of the total amount of samples).

The results are plotted in Figure 8 and Table 1. Notably, the performances on the text-only subset are slightly higher than that on the whole multi-modal IRLBench. For example, o4-mini achieves 77.4% on text-only samples, compared to 76.2% overall performance. Furthermore, results on text-only benchmark also show a clear gap between closed-source and open-source models, in supporting the extremely low-resource Irish language.

5 Statistical Summary

5.1 Subject-Level Overview

Figures 9 average the accuracy of every model within each subject, producing a clean subject-level comparison in both English and Irish. The result shows that Maths and Art both exceed 80 % accuracy, whereas Accounting barely reaches 17 %.

5.2 Language Penalty

Figure 10 also quantifies the cost of switching languages by calculating the mean score of all models in each subject in English and Irish. In this result, scores fall by an average of 18.2 % when the questions are presented in Irish; Art suffers the steepest drop, losing 31.7 %.

5.3 Score Distribution Across Models

Figure 11 treats each model in each subject in each language triplet as a separate data point, so its histogram and box plots capture the full spread of individual scores. The result shows the overall

```
{problem}
Your response should be in the following format:
Answer: <your answer to the above problem>
Confidence: <your confidence score between 0% and 100% for your answer>
```

Figure 1: Prompt used for LLMs to generate response and confidence score.

```
Judge whether the following [response] to [question] is correct or not based on the suggested marking scheme [marking_scheme] below.

[question]: {question} (also in attached images)

[response]: {response}

[marking_scheme]: {marking_scheme} (also in attached images)

Your judgement must be in the format and criteria specified below:

[extracted_final_answer]: The final exact answer extracted from the [response]. Put the extracted answer as 'None' if there is no exact, final answer to extract from the response.

[reasoning]: Explain why the extracted_final_answer is correct or incorrect based on [marking_scheme], focusing only on if the extracted_final_answer follows the [marking_scheme]. Do not comment on any background to the problem, do not attempt to solve the problem.

[correct]: Answer 'yes' if extracted_final_answer follows perfectly the [marking_scheme] given above, or is within a small margin of error for numerical problems. Answer 'no' otherwise, i.e. if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect.

[confidence]: The extracted confidence score between 0% and 100% from [response]. Put 100 if there is no confidence score available.
```

Figure 2: Prompt used for LLM-as-a-judge.

Model	Split	Science	Applied Science	Business Studies	Social Studies	Average
aya-vision-8b	English	33.35	36.29	46.25	56.25	39.37
aya-vision-32b		44.11	45.53	58.96	70.00	50.70
Qwen2.5-VL-32B-Instruct		66.40	63.81	71.72	72.50	67.71
gemma-3-27b-it		65.18	60.30	70.65	71.25	66.77
DeepSeek-R1-70B		67.29	64.82	62.72	74.99	66.46
Llama-4-Scout-Instruct		61.87	57.03	70.71	74.99	64.37
claude-3.7-sonnet		71.25	63.86	75.49	75.00	71.67
gemini-2.0-flash		64.92	60.28	72.09	<u>76.24</u>	67.28
o4-mini		76.73	72.70	82.40	77.50	77.40
gpt-4.1		<u>72.95</u>	<u>66.94</u>	<u>79.58</u>	66.25	<u>72.60</u>
aya-vision-8b	Irish	0.24	0.00	0.98	1.25	0.45
aya-vision-32b		1.42	1.26	0.98	8.75	2.41
Qwen2.5-VL-32B-Instruct		11.87	8.86	3.92	20.14	11.34
gemma-3-27b-it		15.92	14.72	18.63	11.25	15.87
DeepSeek-R1-70B		32.94	30.54	45.10	55.00	36.90
Llama-4-Scout-Instruct		35.00	36.28	54.90	50.00	40.32
claude-3.7-sonnet		40.13	37.66	61.76	<u>66.27</u>	47.19
gemini-2.0-flash		47.63	38.08	<u>66.66</u>	<u>66.25</u>	50.75
o4-mini		<u>51.47</u>	<u>46.86</u>	<u>66.67</u>	<u>66.25</u>	<u>56.35</u>
gpt-4.1		54.18	51.05	67.65	68.75	57.78

Table 1: Model performances per subject group and language on the text-only subset of IRLBench.

```

I will give you an exam paper and the corresponding marking scheme from the official Leaving Certificate Exam of Ireland, containing several problems written in either English or Irish. They are extracted from PDF files as images.
Your task is to extract each problem and the corresponding marking scheme/answer.

Here are some guidelines you should follow
- For each problem found, use the following format:
Problem 1: <problem statement>
Answer 1: <answer to problem 1>

Problem 2: <problem statement>
Answer 2: <answer to problem 2>

...
- For each problem you identify, make sure to keep the original content, including the equations in LaTeX. Remove any redundant context, personal commentary, anecdotes, or unrelated information. But make sure not to change the meaning of the problem and keep all necessary mathematical or technical details.
- For each problem you identify, find the corresponding marking scheme/answer based on question number/part.
- If multiple problems that you extract are related, make sure to include all the context in each problem statement as they will be looked at independently.

Here are a few examples.
{examples}

Please analyze the following exam and extract all math problems. Here are the guidelines one more time for your reference
- For each problem found, use the following format:
Problem 1: <problem statement>
Answer 1: <answer to problem 1>

Problem 2: <problem statement>
Answer 2: <answer to problem 2>

...
- For each problem you identify, make sure to keep the original content, including the equations in LaTeX. Remove any redundant context, personal commentary, anecdotes, or unrelated information. But make sure not to change the meaning of the problem and keep all necessary mathematical or technical details.
- For each problem you identify, find the corresponding marking scheme/answer based on question number/part.
- If multiple problems that you extract are related, make sure to include all the context in each problem statement as they will be looked at independently.

Output:

```

Figure 3: Prompt used for Vision-Language Model Extraction.

mean hovers around 54 %, yet the scatter is wide, especially for the smaller open-source Aya models.

5.4 Model–Subject Interaction

For Figure 12 we compute the mean of the English and Irish accuracies for every cell, letting the heat map place models against subjects. It can be seen that the grid reveals exactly where each model rises or dips across the curriculum.

6 Limitations and Societal Impact

Our benchmark currently addresses a single extremely low-resource language scenario (Irish). Furthermore, the current evaluation methodology leverages an LLM-as-a-judge paradigm, which may present challenges in scalability and robustness. Despite these limitations, the clear performance disparities identified by IRLBench highlight critical gaps in current LLMs, emphasizing the importance of developing robust multilingual generative capabilities. By releasing IRLBench, we hope to foster future research towards improving model performance and linguistic diversity in natural language processing.

References

Example LLM-as-Judge Decision: Correct Answer

<problem>

Is s.a. é an soláthar príomhlíonra cumhacta in Éirinn agus feidhmíonn sé ar mhinicíocht 50 Hz. Biodh sin mar atá, is é s.d. a bhíonn ag teastáil le go leor gléasanna leictreacha sa teach a úsáid.

Idirdhealaigh idir s.a. agus s.d.

<marking_scheme>

athraíonn s.a. a threo
sreabann s.d. in aon treo amháin

<response>

Answer:

S.a. (sruth ailtéarnach) is alternating current (AC), and s.d. (sruth díreach) is direct current (DC).

Sruth ailtéarnach (s.a.): Athraíonn an treo srutha agus an voltas go rialta—athraíonn sé idir dearfach agus diúltach, de ghnáth ag 50 Hz in Éirinn. Úsáidtear sa phríomhlíonra leictreachais, mar atá sna soicéid sa teach.

Sruth díreach (s.d.): Tá an sruth agus an voltas i dtreo amháin i gcónaí—ní athraíonn siad treo. Tagann sé ó bhatairí, painéis gréine, nó asraonta USB agus úsáidtear i go leor feistí leictreacha cosúil le fón póca nó ríomhaire glúine.

Confidence: 100%

<judgement>

{

""extracted_final_answer"": ""S.a. (sruth ailtéarnach) is alternating current (AC), and s.d. (sruth díreach) is direct current (DC). Sruth ailtéarnach (s.a.): Athraíonn an treo srutha agus an voltas go rialta—athraíonn sé idir dearfach agus diúltach, de ghnáth ag 50 Hz in Éirinn. Úsáidtear sa phríomhlíonra leictreachais, mar atá sna soicéid sa teach. Sruth díreach (s.d.): Tá an sruth agus an voltas i dtreo amháin i gcónaí—ní athraíonn siad treo. Tagann sé ó bhatairí, painéis gréine, nó asraonta USB agus úsáidtear i go leor feistí leictreacha cosúil le fón póca nó ríomhaire glúine."",

""reasoning"": ""The response states that s.a. (sruth ailtéarnach) changes direction ('Athraíonn an treo srutha') and that s.d. (sruth díreach) is always in one direction ('Tá an sruth ... i dtreo amháin i gcónaí—ní athraíonn siad treo'). These statements directly match the two criteria in the marking scheme: 'athraíonn s.a. a threo' and 'sreabann s.d. in aon treo amháin.'"",

""correct"": ""**yes**"""

""confidence"": ""100%""

}

Figure 4: Example LLM as Judge Decision: Correct Answer

Example LLM-as-Judge Decision: Incorrect Answer

```

<problem>
Mar bharóin a rangaítear na neodróin a bhíonn páirteach i mbéite-mheath. Mar leaptóin a rangaítear
leictreoin agus posatróin
Tarlaíonn díothú disé nuair a bhuaileann leictreon agus posatrón le chéile. Scríobh cothromóid a léiríonn an
próiseas seo.

<marking_scheme>
$^0_{-1}e \rightarrow ^0_{+1}e \rightarrow 2\gamma
// $2mc^2 = 2hf

<response>
Answer: \(\ e^- + e^+ \rightarrow 2\gamma \)
Confidence: 100%

<judgement>
{
  ""extracted_final_answer"": ""e^- + e^+ \rightarrow 2\gamma\nThe equation e- + e+ -> 2\text{gamma} is presented
  as the answer."",
  ""reasoning"": ""The response provides the equation $e^- + e^+ \rightarrow 2\text{gamma}$ for electron-
  positron annihilation. The marking scheme shows $^0_{-1}e \rightarrow ^0_{+1}e \rightarrow 2
  \text{gamma}$. The notation for the particles ($e^-, e^+$ vs $^0_{-1}e, ^0_{+1}e$) and the operators
  separating them (+ vs $\rightarrow$) do not match the marking scheme."",
  ""correct"": ""no"",
  ""confidence"": ""100%"""
}

```

Figure 5: Example LLM as Judge Decision: Incorrect Answer

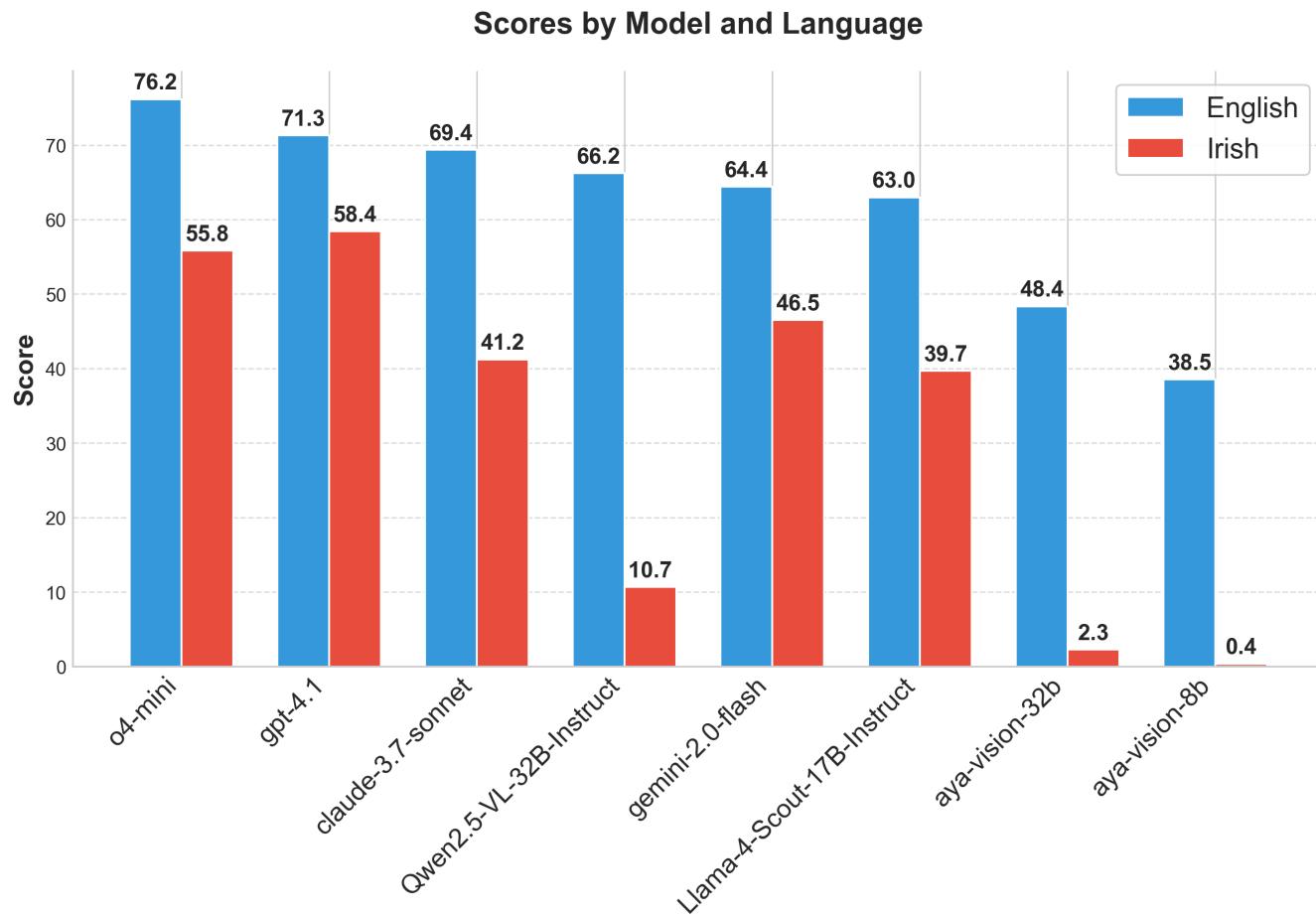


Figure 6: Accuracy scores on IRLBench per model and language with additional models.

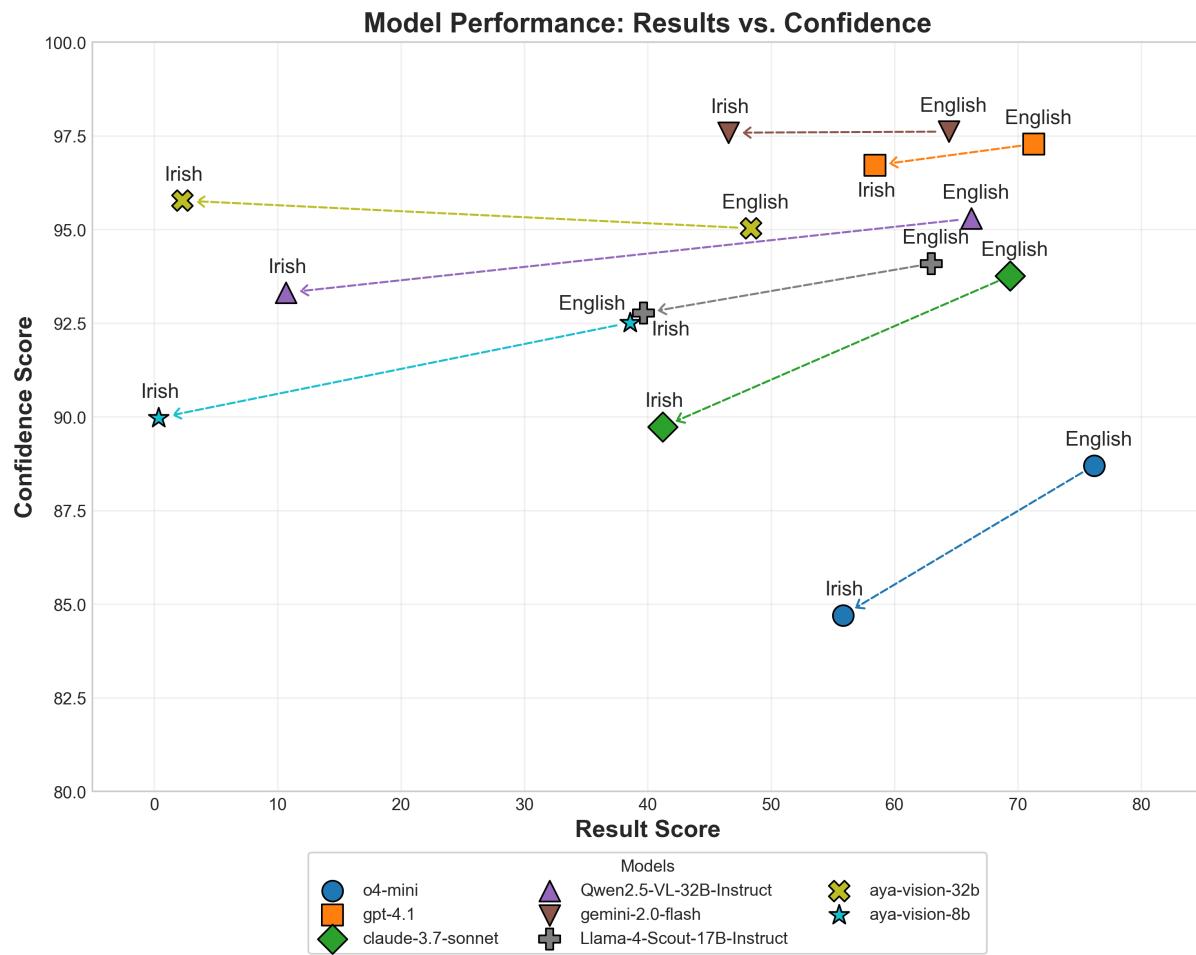


Figure 7: Model self-reported confidences compared to performances.

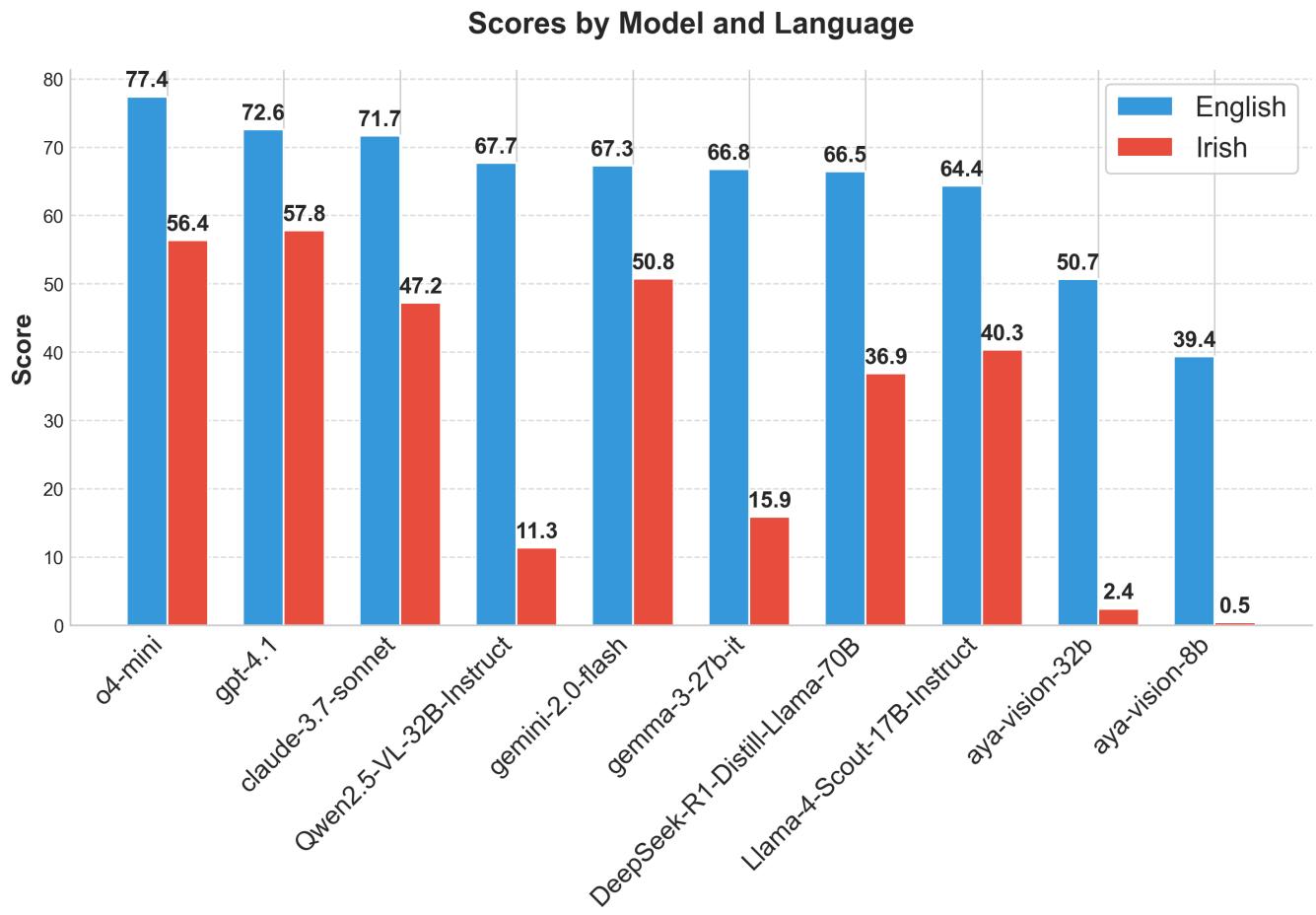


Figure 8: Accuracy scores on text-only subset of IRLBench per model and language.

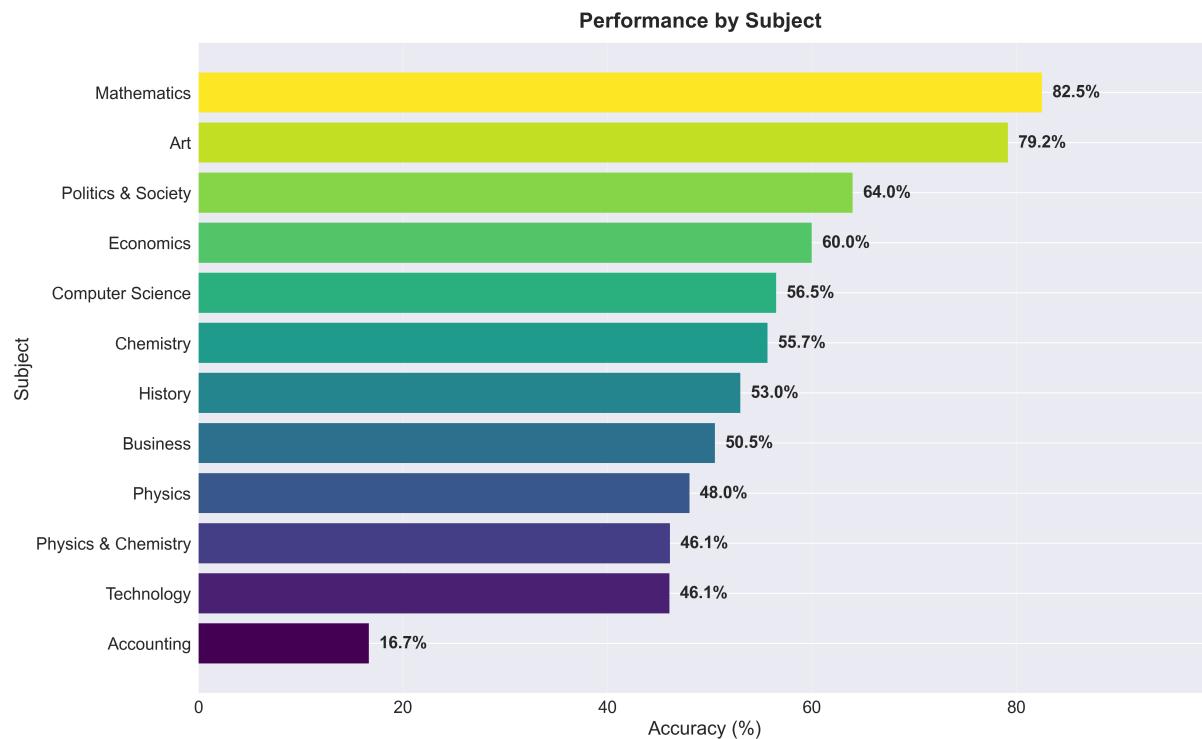


Figure 9: Accuracy Breakdown by Subject.

English vs Irish Language Analysis

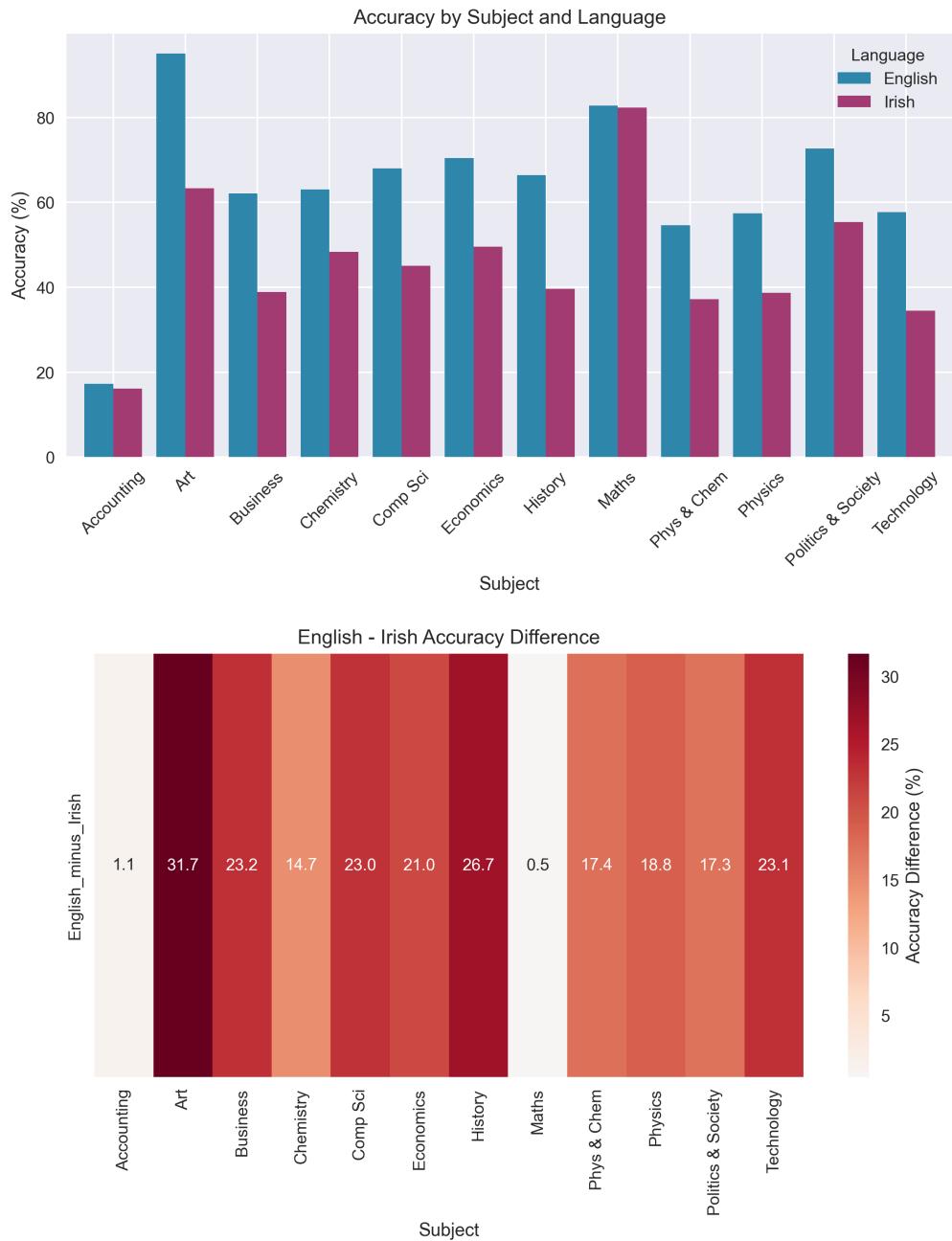


Figure 10: Accuracy Breakdown by Language

Statistical Analysis of Accuracy Distributions

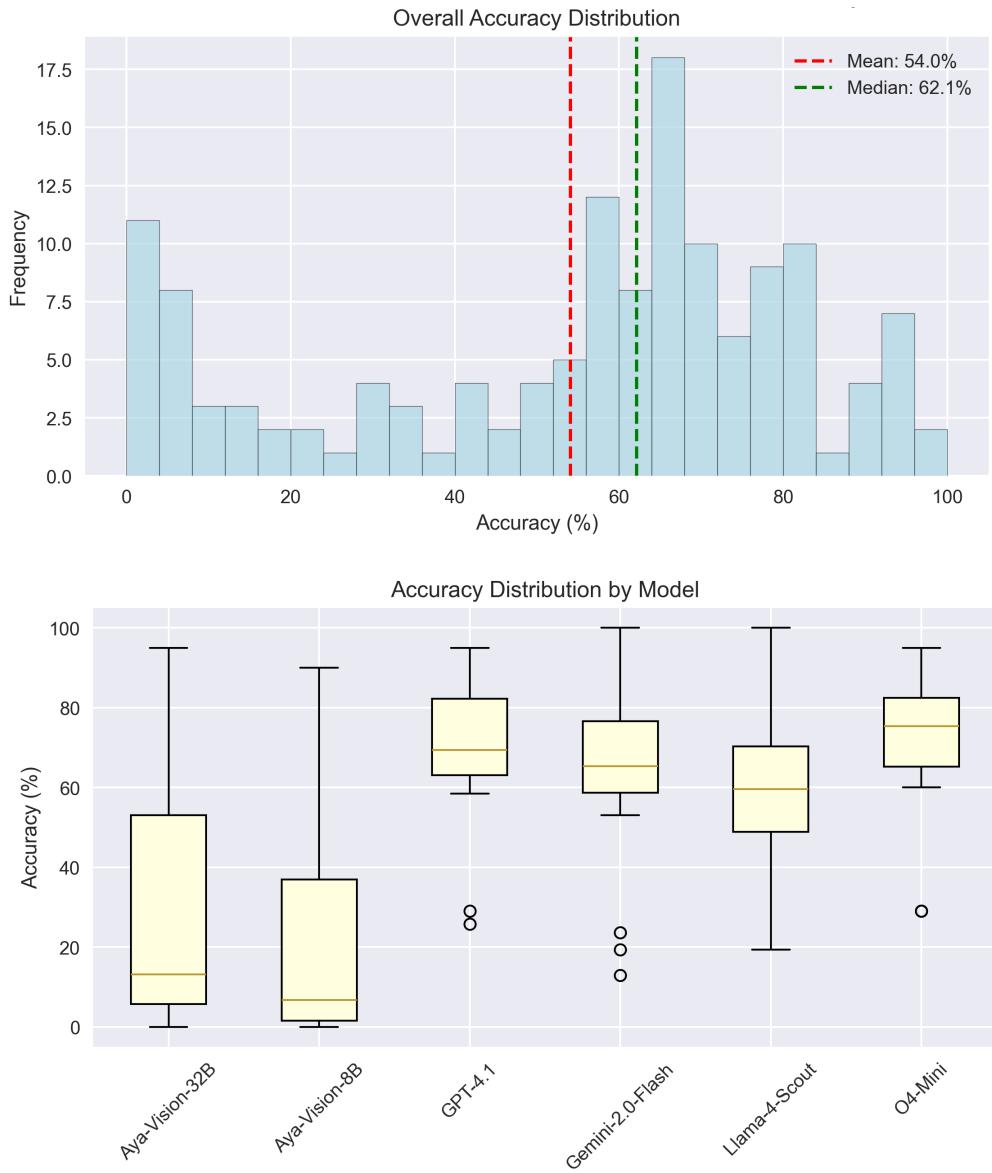
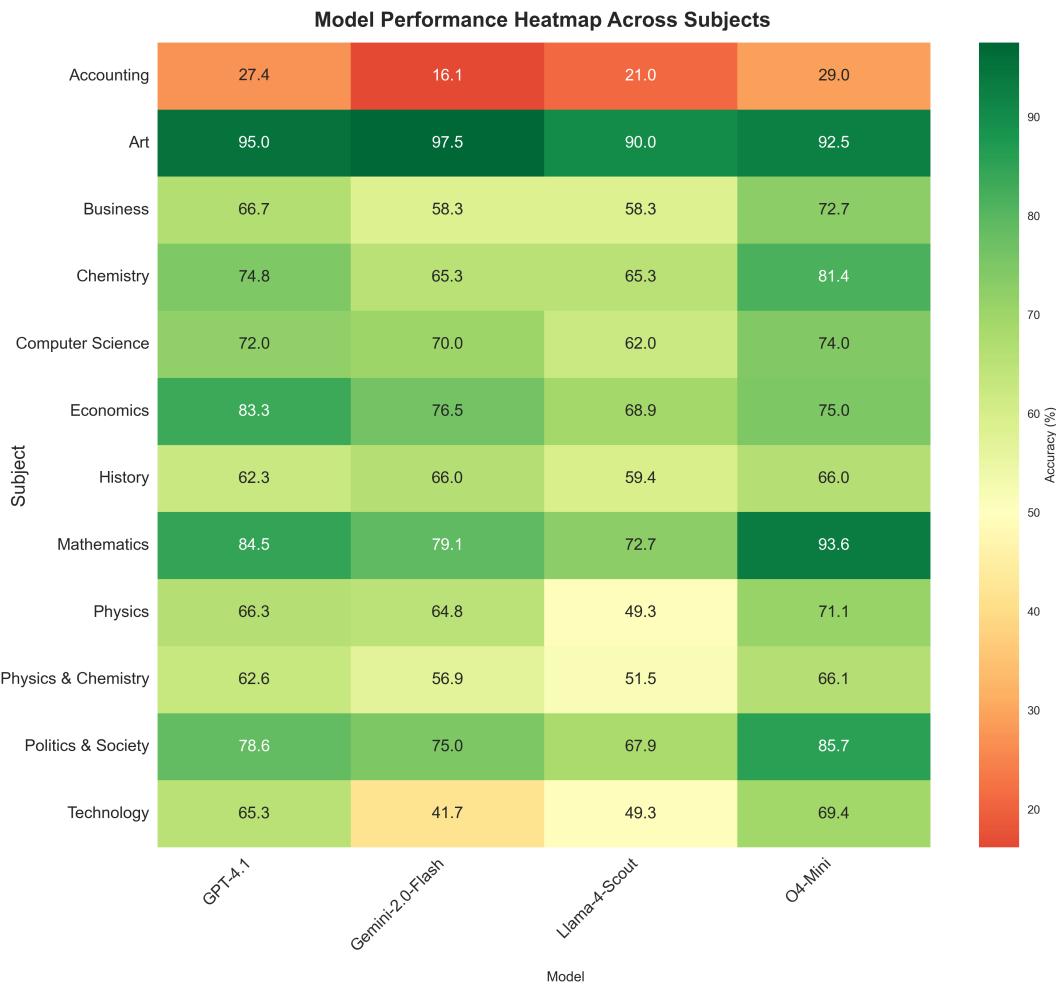


Figure 11: Statistical analysis of Accuracy distributions

**Figure 12: Model Performance Heatmap across subjects**